Hierarchical AI-Meteorologist: LLM-Agent System for Multi-Scale and Explainable Weather Forecast Reporting

Daniil Sukhorukov[1, 2], Andrei Zakharov[1], Nikita Glazkov[1, 3], Katsiaryna Yanchanka[4], Vladimir Kirilin[4], Maxim Dubovitsky[4], Roman Sultimov[5], Yuri Maksimov[6], Ilya Makarov[1, 7, 8]

[1]AI Research Institute [2]ITMO University [3]NUST MISIS [4]Independent Researcher [5]AI Center, Lomonosov Moscow State University [6]LLC Interdata [7]ISP RAS [8]Research Center of the Artificial Intelligence Institute, Innopolis University

Abstract

We present the Hierarchical AI-Meteorologist, an LLM-agent system that generates explainable weather reports using a hierarchical forecast reasoning and weather keyword generation. Unlike standard approaches that treat forecasts as flat time series, our framework performs multi-scale reasoning across hourly, 6-hour, and daily aggregations to capture both short-term dynamics and long-term trends. Its core reasoning agent converts structured meteorological inputs into coherent narratives while simultaneously extracting a few keywords effectively summarizing the dominant meteorological events. These keywords serve as semantic anchors for validating consistency, temporal coherence and factual alignment of the generated reports. Using OpenWeather and Meteostat data, we demonstrate that hierarchical context and keyword-based validation substantially improve interpretability and robustness of LLM-generated weather narratives, offering a reproducible framework for semantic evaluation of automated meteorological reporting and advancing agent-based scientific reasoning.

1 Introduction

Automating the interpretation of tabular, hourly weather forecasts for specific locations remains a nontrivial challenge at the intersection of meteorology and data-to-text generation. Prior works in weather Natural Language Generation (NLG), including the SumTime projects, established the importance of content selection and lexical choice when translating multivariate time series into coherent text conclusions and provided parallel "data $\leftrightarrow$ description" corpora (Reiter et al. 2005; Belz 2005, 2008). In operational practice, National Weather Service forecasters' Area Forecast Discussions (AFD) serve as reference texts that articulate causal reasoning and confidence levels (National Weather Service 2010, 2024). Yet a gap persists between dense numerical tables and human-readable reports, where causal links and verifiability are critical.

Despite substantial progress in machine-learning-based weather prediction, exemplified by recent systems targeting medium-range horizons (Lam et al. 2023; Bi et al. 2023; Rasp et al. 2024; Shi et al. 2025), a practical question arises: how can their tabular outputs be transformed into explainable and verifiable textual reports? In this paper we treat such models purely as sources of forecast tables and intentionally avoid the modeling details. Instead, our focus is on the interpretation of these weather forecasts. Even applying LLMs and VLMs to meteorological tasks, from imagery interpretation to risk communication (Lawson et al. 2025; Franch et al. 2024; Jin et al. 2023; Chen et al. 2025), are not able to provide multi-scale interpretations of numerical data across several time resolutions at once. The need in interpretation is especially relevant for medium-range horizons beyond five days, where large tables of detailed data confound local fluctuations with dayly trends, requiring causal interpretability and readability.

In this work, we introduce the Hierarchical AI-Meteorologist, an LLM-agent pipeline that performs hierarchical interpretation of forecast tables at three concurrent levels: hourly (local dynamics), six-hourly (mesoscale patterns and noise smoothing), and daily (persistent trends and synoptic transitions). The resulting report is organized as a single narrative with consistent events and explanations across different levels. A key element of the proposed framework is the synthesis of weather keywords, a compact set of three to five terms or phrases that summarize dominant weather states and their evolution for the target time interval. For extraction and consistency control, we adapt robust keyword/keyphrase methods to the meteorological domain (Mihalcea and Tarau 2004). A second element is the proof-block, a brief structured "evidential" insert enumerating table-derived signals (pressure tendencies, wind shifts and strengthening, daily temperature amplitudes, precipitation duration and intensity) that support each generated keyword. This coupling supplies semantic anchors and verifiability since a stated event must manifest in observable aggregates and patterns.

The system operates on open data sources: hourly forecast tables from OpenWeather (One Call 3.0) and climatological background from Meteostat (monthly aggregates and climate normals), enabling scalable application to diverse locations without model fine-tuning (OpenWeather 2025; Meteostat 2021, 2022). For the United States we additionally consult AFD as a weak consistency reference without requiring textual overlap (National Weather Service 2010, 2024). On the LLM side, we rely on in-context serialization of nu-

merical data and prompts rather than task-specific training, lowering deployment barriers and improving reproducibility.

This work makes the following key contributions:

- a hierarchical scheme for interpreting tabular forecasts (hourly → six-hourly → daily) and composing a causally consistent cross-scale narrative;
- the introduction of weather keywords as a semantic validation layer and concise summary, linked to the data through a structured proof-block;
- a practical, reproducible integration of open sources (OpenWeather, Meteostat) and operational texts (AFD) into an in-context LLM interpretation pipeline.

2 Related Work

LLM-agent systems for scientific reasoning. Recent work has demonstrated the emergence of LLM-agent systems as a new paradigm for scientific reasoning. Early systems such as AutoGPT (Significant-Gravitas 2023) and MetaGPT (Hong et al. 2023) introduced autonomous multi-agent collaboration frameworks, while AutoGen (Wu et al. 2024) formalized conversational agent orchestration for complex analytical tasks. In scientific domains, ChemCrow (M. Bran et al. 2024) and AI Scientist (Lu et al. 2024) demonstrated how LLM agents can autonomously design experiments, retrieve literature, and validate hypotheses. These systems collectively illustrate the growing shift from static language models to interactive, tool-augmented scientific agents capable of structured reasoning and knowledge generation.

LLMs in meteorology and weather forecasting. Recent work has begun leveraging LLMs to translate structured meteorological data into human-interpretable weather narratives and interactive tools. For example, ECMWF's DestinE chatbot (ECMWF 2025) to make high-resolution weather and climate data accessible via conversational interfaces, CLLMate (Li et al. 2024) to enable event-based forecasting of weather and climate phenomena in text form, and GPT-based forecasting (Franch et al. 2025) to perform precipitation nowcasting by tokenizing radar imagery and generating ensemble forecasts via language-model driven frameworks. More recent works have explored LLM-agent frameworks that integrate iterative querying, reflection, and domain-specific validation to improve the scientific accuracy and robustness of weather reports (Varambally et al. 2025).

3 Methodology

Data acquisition. All inputs are assembled by the non-LLM *Assistant* block. Location metadata includes city, administrative region, country, elevation above mean sea level, and an optional short description from Wikipedia when available. Climatological context (an example depicted in Fig.3 in SM) is retrieved primarily from Meteostat (monthly aggregates and climate normals); when Meteostat is unavailable, we fall back to an ERA5-based monthly climatology on a 0.25° grid with nearest-neighbor extraction for the requested point. For each location there are

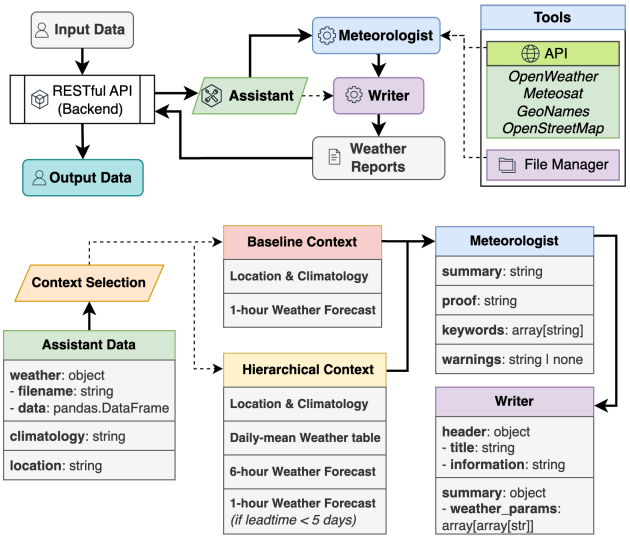

Figure 1: Overall architecture of the Hierarchical AI-Meteorologist combining three key blocks (Assistant, Meteorologist, Writer) to automatically generate coherent weather reports.

monthly $\{\mathrm{T_{min}}, \mathrm{T_{max}}, \mathrm{P_{tot}}\}$ (minimum/maximum air temperature and total precipitation). Hourly forecasts are obtained from OpenWeather One Call (time grid $\Delta t = 1\,\mathrm{h}$) and include forecast timestamp, categorical weather condition, near-surface air temperature T [°C], feeling temperature T_{feel} [°C], dew point T_d [°C], relative humidity RH [%], wind speed U [$\mathrm{m\,s^{-1}}$], wind direction θ [deg], wind gust G [$\mathrm{m\,s^{-1}}$], liquid/solid precipitation amount P [mm], and horizontal visibility Vis [m]. To improve textual outputs downstream, a compact weather category is also assigned via rule-based thresholds consistent with OpenWeather condition codes.

Hierarchical temporal aggregation. To support multi-scale interpretation while controlling context length, the Assistant block forms two groups of aggregates over non-overlapping windows $W \in \{6\mathrm{h}, 1\mathrm{d}\}$:

$$\overline{T}_W = \frac{1}{|W|}\sum_{t\in W} T_t, \quad T_W^{\max} = \max_{t\in W} T_t, \quad T_W^{\min} = \min_{t\in W} T_t,$$

$$\overline{\mathrm{RH}}_W = \frac{1}{|W|}\sum_{t\in W} \mathrm{RH}_t, \quad \overline{U}_W = \frac{1}{|W|}\sum_{t\in W} U_t, P_W = \sum_{t\in W} P_t.$$

Wind direction is averaged on the circle,

$$\bar{\theta}_W = atan2\Big(\frac{1}{|W|}\sum_{t\in W}\sin\theta_t,\ \frac{1}{|W|}\sum_{t\in W}\cos\theta_t\Big)\ [^\circ],$$

and is stored for 6-hour windows; for daily windows we omit $\bar{\theta}_{1\mathrm{d}}$ to avoid misleading circular averages. Dew point and visibility are aggregated as means; winds are summarized by mean and (optionally) maxima. Thus each 6-hour record contains $\{\,\overline{T}, T^{\max}, T^{\min}, \overline{\mathrm{RH}}, \overline{U}, \bar{\theta}, P\,\}$; each daily record mirrors this set except for wind direction.

Report formation and output structure. The *Meteorologist* agent uses available context and generates a structured analysis with four fields: `summary` (multi-paragraph narrative grounded in the supplied tables), `proof` (a compact evidential rationale that points to table-derived patterns such as pressure tendencies, wind shifts/strengthening, daily temperature amplitudes, and precipitation duration/intensity), `keywords` (a list of 3-5 weather descriptors summarizing dominant states and transitions over the forecast horizon), and optional `warnings` (flagging anomalous or hazardous conditions). The downstream *Writer* agent adapts this analysis to the user's domain and style preferences and returns a JSON report over REST with a minimal, explicit format. A typical response includes

- `header:{title, information}` for location-aware titling and a brief preamble,
- `analysis:{summary, proof, keywords, warnings}`,
- `context:{mode, daily, six_hour, hourly, climatology, location}` showing the actually used tables.

This representation makes the narrative consistent because every declared keyword is expected to be supported by at least one entry in the proof block and by observable aggregates in the corresponding tables.

Next, we describe the proposed framework in terms of the data-processing pipeline, the construction of multi-level context, and the two-step reasoning procedure (Meteorologist $\rightarrow$ Writer). Illustrative diagram of the overall architecture and of context flows between agents are provided in the Figure 1 with the main blocks and agent links.

3.1 System overview

The pipeline follows a microservice (RESTful) paradigm. The external-data processing block *Assistant* collects and normalizes location-specific inputs across meteorology, climatology, and geodata, then builds aggregates (6-hour and daily windows), and packages one of two context modes (baseline or hierarchical). The *Meteorologist* block is an LLM agent with a structured output (`summary`, `proof`, `keywords`, `warnings`) that interprets tabular data and captures causal relations. The *Writer* block is an LLM agent for post-editing to the target domain style and report format; it does not alter the factual basis produced by the Meteorologist agent, but adapts it to the user's stylistic request, returning a JSON report together with the context, targeted to the user domain (risk analysis, energy, extreme weather).

3.2 Input collection and normalization (Assistant)

Location. We extract city name, administrative region, country, elevation above the mean sea level. When available, a concise Wikipedia description is included in the report preamble.

Climatology. The primary source is Meteostat (monthly aggregates and normals). When services are unavailable, the system returns back to monthly ERA5 climatology on a 0.25° grid with nearest-neighbor extraction at the query point. For each month we store $\{\mathrm{T_{min}}, \mathrm{T_{max}}, \mathrm{P_{tot}}\}$.

Hourly forecast. From OpenWeather (One Call 2.5) we use table data with $\Delta t = 1\,\mathrm{h}$ consisting a timestamp, categorical "weather icon/state", T (air temperature), T_{feel} (feels-like), T_d (dew point), RH (relative humidity), U (wind speed), θ (wind direction), G (gust), P (precipitation), and Vis (visibility). To enhance textual outputs we additionally assign a compact weather category based on threshold rules consistent with OpenWeather codes, i.e., derived from thresholded assessments of the meteorological variables.

3.3 Multi-level aggregation and context modes

To enable interpretation at multiple time scales while controlling context length, the *Assistant* aggregates over non-overlapping windows $W \in \{6\mathrm{h}, 1\mathrm{d}\}$ using means/minima/-maxima for T, means for RH and U, sums for P. Wind direction is averaged on the circle and stored only for 6-hour windows. The resulting modes are:

- **Baseline Context**: location and climatology plus the full hourly table (short-range forecasts, no hierarchy).
- **Hierarchical Context**: location and climatology, a daily table, and a 6-hour aggregate table for the location; when the lead time $H < 5$ days, the hourly table is additionally included. For $H \in [5, 10]$ days the hourly grid is omitted to save tokens and reduce LLM attention bias toward large tables.

Both modes are serialized into a single payload and cached (replay without repeated external API calls; ability to run different "meteorologists" over the same frozen sample).

3.4 Stage 1: Interpreting tabular data (Meteorologist)

Output schema. The agent uses available context and returns a structured analysis consisting:

- `summary` — several paragraphs describing the weather dynamics across the horizon, based on the supplied tables (daily/6h/1h);
- `proof` — a compact block listing observable patterns (pressure tendencies, wind shifts/strengthening, daily amplitude of T, duration/intensity of precipitation, etc.) that explain the observed events; the goal is to make the report verifiable and explainable for meteorological experts;
- `keywords` — 3–5 *weather keywords* summarizing dominant weather states and transitions over the time interval (e.g., "cooling; brief rain; wind strengthening"). Generation is performed by the LLM with a controlled vocabulary and guidance to rely on aggregates; each keyword is expected to correspond to at least one feature in the `proof`;
- `warnings` — optional anomalous/hazardous phenomena (strong winds, intense precipitation, icing, etc.) with brief data-grounded justification, when the model flags expected hazard.

3.5 Stage 2: Domain/style adaptation (Writer)

The *Writer* receives the structured analysis and user parameters (tone, length, application domain) and composes a report with a predefined JSON structure consisting:

- `header:{title, information}` — a location-aware title and short preamble;
- `analysis:{summary, proof, keywords, warnings?}` — the substantive content forwarded from the *Meteorologist* with minimal stylistic edits;
- `context:{mode, daily, six_hour, hourly?, climatology, location}` — an echo block listing the tables actually used to generate the text.

The Writer does not change facts from the Meteorologist. Its primary role is to adapt exposition and layout to the target user (e.g., power engineer, urban planner, agronomist).

3.6 RESTful integration and reproducibility

The part is split into two components: *Analysis* and *Report*. *Analysis* accepts the serialized context and returns the structured output from the *Meteorologist*. *Report* composes the final report from the *Writer*. Caching of "raw" and aggregated contexts (OpenWeather, Meteostat/ERA5) enables reproducing reports when models and prompt templates change, without repeated network calls. Typical reliability elements include retries for external APIs, explicit degradation codes (e.g., fallback to ERA5 climatology), validation of input JSON, and completeness checks for required fields prior to passing data to LLM agents.

3.7 Current limitations

The system does not perform separate "tabular reasoning" with programmatic hypothesis testing; verifiability is realized through the two-step textual rationale (`summary` + `proof`) and its linkage to `keywords`. In the present version, reports are returned as JSON, and interfaces are provided for attaching agent-based components for PDF/plot generation via preconfigured environments and Python libraries.

4 Results

We demonstrate the system performance on four different locations of distinct climate and weather behavior: Cork, Ireland (51.903614° N, -8.468399° E), Manila, Philippines (14.5995° N, 120.9842° E), Chennai, India (13.0827° N, 80.2707° E), and Hai Châu, Da Nang, Vietnam (16.0472° N, 108.2200° E). In all cases the forecast horizon is $\sim$5–6 days (120 hours) in late October, 2025. Reports are generated in the hierarchical mode (daily + 6h; hourly rows added for short sub-intervals). We evaluate three key characteristics of report quality: (i) consistency of the `summary` with tabular aggregates, (ii) alignment of `keywords` with observed patterns, and (iii) adequacy of `proof`/`warnings` relative to the given data and climatology.

Cork, Ireland (fall transition in a mild coastal climate). The summary (Fig.4a in SM) effectively captures a smooth cooling trend within forecast: daytime maxima decrease from $\approx$14.4–15.5°C early in the window to $\approx$10–11.7°C by October, 23–24; relative humidity often exceeds 80%; precipitation is intermittent and light (daily total up to 5.4 mm on October, 23). Winds are moderate (2–7.8 $\mathrm{m\,s^{-1}}$) with a shift from S to W–NW, interpreted as passage of a weak frontal disturbance. The generated keywords `keywords` = *cooling_trend, light_rain, moist_conditions, frontal_passage, autumn_transition* compactly reflect the dominant evolution and are supported by aggregates of T (declining maxima), RH (elevated means), P (small totals), and the wind-direction shift. The narrative is readable with no false alarms; `warnings` did not trigger.

Manila, Philippines (persistent tropical regime with coastal influence). The report (Fig.4b in SM) describes identified climatologically typical warm and humid conditions: maxima $\approx$27.6–31.8°C, minima $\approx$25.4–26.9°C; humidity 66–90%; precipitation infrequent and light (daily usually < 5 mm); winds light to moderate, predominantly E–SE, commonly < 4 $\mathrm{m\,s^{-1}}$. `keywords` = *light_rain, stable_conditions, marine_influence, warm_anomaly, clear_sky* are consistent with small rainfall totals, weak winds, and a coastal regime; a local ambiguity may arise for *warm_anomaly* when maxima are slightly below the climatological mean—this highlights sensitivity of the keyword to the chosen reference and thresholds. Otherwise, agreement between daily/6-hour aggregates and the `summary` is stable. `warnings` are not flagged, which perfectly corresponds to a non-extreme scenario.

Chennai, India (transition from humid/windy to cooler and drier). The data in report ((Fig.4c in SM)) show gradual cooling and a reduction in rainfall toward the end of the window: maxima fall from $\approx$30.7°C (October, 21) to $\approx$25.3°C (October, 24), minima from $\approx$26.6°C to $\approx$23.4°C; humidity remains high (73–92%). Wind speeds increase from $\approx$3.3 to > 10 $\mathrm{m\,s^{-1}}$ by October, 24 with direction shifting from E toward NW, interpreted as influence of a frontal-like process. Rainfall is substantial early (up to 27.7 mm daily on October, 20) with subsequent tapering. The system generates the following keywords `keywords` = *heavy_rain, frontal_passage, strong_wind, overcast, stable_conditions* capturing a transition regime in the weather ("humid and windy" $\rightarrow$ "cooler and drier"); the `proof` block correctly relates these tags to elevated early P totals, rising U and a θ shift, and to decreasing daily T maxima.

Hai Châu (Da Nang, Vietnam; intense rainfall and wind with active warnings). The summary in Fig.2 describes persistently high humidity and frequent precipitation, with an extreme daily total on October, 23 $P > 130$ mm and increased wind speeds up to 9.2 $\mathrm{m\,s^{-1}}$. A strong wind-direction shift (tens of degrees) is observed along with atmosphere temperatures "below the climatological maximum" for October, followed by a gradual decrease of rainfall and wind to October, 25. The generated `warnings` field is triggered, providing a valuable information about a risk of lo-

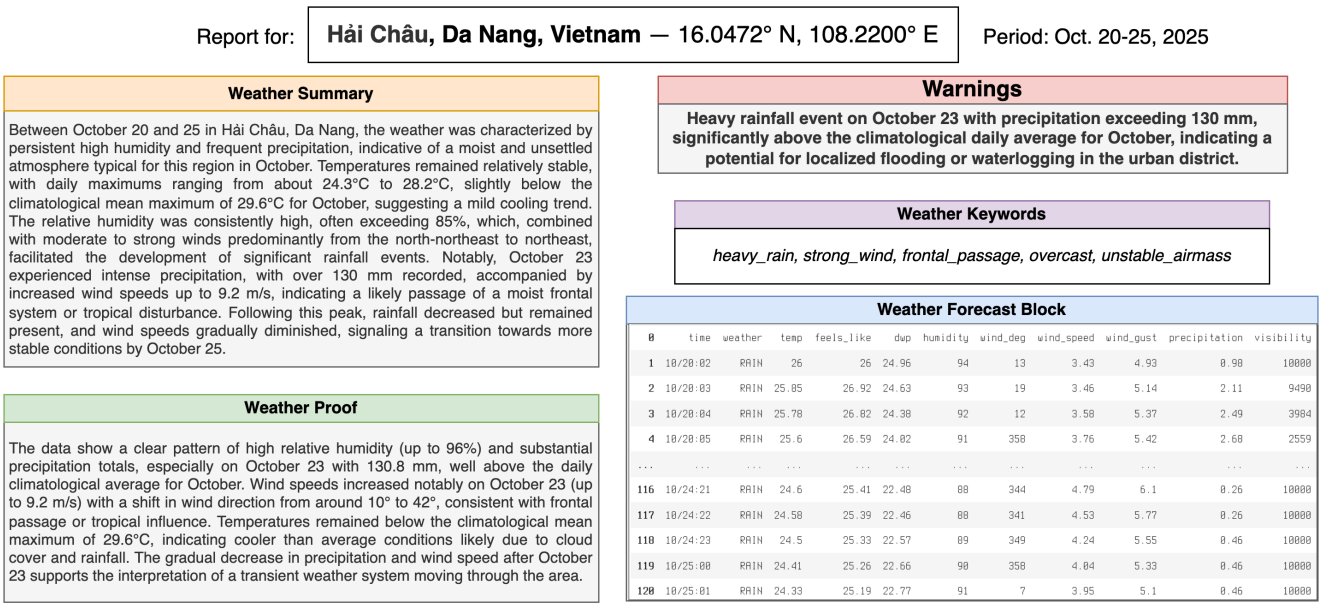

Figure 2: An example of generated report by reasoning on forecasting data with extreme events. The system automatically provides a brief summary, keywords and warnings for anticipated events.

calized flooding due to increased above the climatological average daily total rain rate. The corresponding keywords are `keywords` = *heavy_rain, strong_wind, frontal_passage, overcast, unstable_airmass*; the `proof` provides information about very high RH (up to 96%), the extreme daily P, strengthening U, and a θ shift, making this case illustrative for validating hazardous events in the tropics.

Comparative analysis and observed model behavior. (1) Hierarchical presentation (daily + 6h) improves narrative coherence: for Cork and Chennai locations, the daily trend in T is clearly separated from intraday variability, while in Da Nang a large daily extreme is not "lost" amid numerous hourly rows. (2) The `keywords`–`proof` coupling simplifies evaluation: for each location, keywords are supported by aggregates (P, U, θ, T, RH); in Manila, borderline values for *warm_anomaly* indicate that thresholds and/or reference normals may require refinement for the tropics. (3) `warnings` in those locations were triggered only in the presence of explicit extreme conditions (Da Nang), consistent with the rule design requiring large deviations from climatology and/or exceeding the predefined P and U thresholds. (4) The content of the `summary` remains aligned with threshold-based weather labels in the forecast tables and no false descriptions are observed, and "frontal" tags (*frontal_passage*) correlate with wind shifts and the phase structure of precipitation.

Overall, these demonstrated cases show that hierarchical context, combined with controlled `keywords` and the evidential `proof` insert, yields readable and verifiable reports for a ∼5–6 day horizon across diverse climate zones, from mild coastal transition patterns to typical tropical regimes in the weather with extreme rainfall events.

5 Future Work

The system can be further improved along three potential directions:

1. **AFD-style benchmark and auto-correction.** A NOAA AFD–inspired corpus pairing forecast tables with forecaster discussions and an error taxonomy (false/missed events, wrong trend sign, keyword–aggregate mismatch, over/under-warning) can sufficiently enhance the model performance. Additionally, a critic–corrector loop will compare `summary`/`keywords`/`proof` with daily/6h/1h aggregates and apply minimal changes (rephrase the trend description, change/remove keywords, improve the warning), exposed as lightweight *Validator*/*Editor* services on the current REST architecture.
2. **Ensemble-aware interpretation.** We will augment the context with distribution tables (mean/median, p10–p90, spread, exceedance fractions) such that reports can include probability-tagged keywords (e.g., *heavy_rain [40–60%]*), enclose trends under large spread, and add graded `warnings`. This opens ways for new experiments with the model, its probabilistic calibration and decision evaluation.
3. **ReAct tooling for targeted validation.** An aggregate-level detector will flag uncertain and dangerous events (extreme P, abrupt wind shifts, keyword–data worry). A ReAct agent (Yao et al. 2022) will then hypothesize, *zoom* to hourly rows for the flagged window, query diagnostics (gradients, run-lengths, gust quantiles) or fallback climatology, and minimally update `proof`/`keywords`/`warnings`. Tools include *aggregate-checker*, *hourly-fetch*, *threshold-tester*, and *climatology-lookup*.

Together, these improvements aim to make the AI-Meteorologist framework self-correcting, ensemble-aware, and able to manage with uncertain intervals on demand while preserving the current microservice/REST architecture.

6 Conclusion

Our experiments show that encoding forecast context at multiple temporal scales and linking concise, data-grounded keywords to an evidential proof block materially improves the usability and verifiability of LLM-generated weather narratives. Across four geographically and climatologically distinct case studies (Cork, Manila, Chennai, Da Nang) we found systm improved performance in (i) narrative–table consistency, (ii) keyword–aggregate alignment, and (iii) the relevance of proof/warning items—most notably, the system flagged and justified a hazardous rainfall/wind episode in Da Nang while avoiding false alarms in non-extreme cases. Importantly for environmental applications, the hierarchical context also reduced token-bias toward noisy hourly rows, helping the system separate daily variability from persistent, decision-relevant trends.

While the developed system already demonstrates strong promise, several clear and readily actionable enhancements can further strengthen its operational value. It can be further improved by complementing the existing text-based verification with programmatic checks, localizing keyword thresholds to improve tropical and regional sensitivity, and incorporating ensemble and uncertainty information to produce probability-aware summaries. Concrete next steps include building an AFD-style benchmark and critic–editor loop, adding ensemble-aware, probability-tagged keywords, and deploying ReAct-style tooling for targeted aggregate checks. These improvements will accelerate the system's capabilities from a reproducible research prototype into an operational, self-checking pipeline for explainable meteorological reporting.

Collectively, this positions our Hierarchical AI-Meteorologist as a reproducible, explainable bridge between numerical forecasts and environmental decision-support. It offers interpretability, scientific rigor, and deployability for stakeholders in climate resilience, emergency management, and resource planning.

7 Acknowledgements

The work of I. Makarov was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

References

Belz, A. 2005. Corpus-driven generation of weather forecasts. In *Proc. 3rd Corpus Linguistics Conference*.

Belz, A. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4): 431–455.

Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970): 533–538.

Chen, J.; Zhou, P.; Hua, Y.; Chong, D.; Cao, M.; Li, Y.; Chen, W.; Zhu, B.; Liang, J.; and Yuan, Z. 2025. ClimateIQA: A New Dataset and Benchmark to Advance Vision-Language Models in Meteorology Anomalies Analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 5322–5333.

ECMWF. 2025. Development Seed to create a climate and weather chatbot in DestinE. "News" article on the Destination Earth website. https://destine.ecmwf.int/news/development-seed-to-create-a-climate-and-weather-chatbot-in-destine/?utm_source=chatgpt.com (accessed: 2025-10-22).

Franch, G.; Tomasi, E.; Wanjari, R.; Poli, V.; Cardinali, C.; Alberoni, P. P.; and Cristoforetti, M. 2024. GPTCast: a weather language model for precipitation nowcasting. *arXiv preprint arXiv:2407.02089*.

Franch, G.; Tomasi, E.; Wanjari, R.; Poli, V.; Cardinali, C.; Alberoni, P. P.; and Cristoforetti, M. 2025. GPTCast: a weather language model for precipitation nowcasting. *Geoscientific Model Development*, 18(16): 5351–5371.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; et al. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. 2023. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421.

Lawson, J. R.; Trujillo-Falcón, J. E.; Schultz, D. M.; Flora, M. L.; Goebbert, K. H.; Lyman, S. N.; Potvin, C. K.; and Stepanek, A. J. 2025. Pixels and predictions: potential of GPT-4V in meteorological imagery analysis and forecast communication. *Artificial Intelligence for the Earth Systems*, 4(1): 240029.

Li, H.; Wang, Z.; Wang, J.; Wang, Y.; Lau, A. K. H.; and Qu, H. 2024. CLLMate: A Multimodal Benchmark for Weather and Climate Events Forecasting. *arXiv preprint arXiv:2409.19058*.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; and Schwaller, P. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5): 525–535.

Meteostat. 2021. Meteostat API: Monthly Data.

Meteostat. 2022. Meteostat API: Climate Normals.

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.

National Weather Service. 2010. The Area Forecast Discussion (AFD). Product purpose and structure.

National Weather Service. 2024. National Weather Service Web API Documentation.

OpenWeather. 2025. OpenWeather API Guide.

Rasp, S.; Hoyer, S.; Merose, A.; Langmore, I.; Battaglia, P.; Russell, T.; Sanchez-Gonzalez, A.; Yang, V.; Carver, R.; Agrawal, S.; et al. 2024. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6): e2023MS004019.

Reiter, E.; Sripada, S.; Hunter, J.; Yu, J.; and Davy, I. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2): 137–169.

Shi, J.; Shirali, A.; Jin, B.; Zhou, S.; Hu, W.; Rangaraj, R.; Wang, S.; Han, J.; Wang, Z.; Lall, U.; et al. 2025. Deep learning and foundation models for weather prediction: A survey. *arXiv preprint arXiv:2501.06907*.

Significant-Gravitas. 2023. AutoGPT. https://github.com/Significant-Gravitas/AutoGPT. Version v0.6.33, accessed: 2025-10-22.

Varambally, S.; Fisher, M.; Thakker, J.; Chen, Y.; Xia, Z.; Jafari, Y.; Niu, R.; Jain, M.; Manivannan, V. V.; Novack, Z.; et al. 2025. Zephyrus: An Agentic Framework for Weather Science. *arXiv preprint arXiv:2510.04017*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Supplementary Material

Location Block

CITY: City of New York # REGION: New York #
COUNTRY: United States
MEAN ELEVATION: 8.0 meter #
LOC DESCRIPTION (APPROXIMATELY): ...

Climatology Block

CLIMATOLOGY MEAN FOR YEARS: [1991-2020]
LABELS: *[TMAX = TEMPERATURE MAX (°C),*
TMIN = TEMPERATURE MIN (°C),
TP = TOTAL PRECIPITATIONS (mm)]

MONTH: 1 # TMAX: 4.6 # TMIN: -3.4 # TP: 87.6
MONTH: 2 # TMAX: 6.2 # TMIN: -2.5 # TP: 75.7
MONTH: 3 # TMAX: 10.7 # TMIN: 1.3 # TP: 104.8
MONTH: 4 # TMAX: 17.1 # TMIN: 6.9 # TP: 98.4
MONTH: 5 # TMAX: 22.8 # TMIN: 12.3 # TP: 100.8
MONTH: 6 # TMAX: 27.9 # TMIN: 17.7 # TP: 110.2
MONTH: 7 # TMAX: 30.7 # TMIN: 21.0 # TP: 118.5
MONTH: 8 # TMAX: 29.5 # TMIN: 20.1 # TP: 105.4
MONTH: 9 # TMAX: 25.6 # TMIN: 16.1 # TP: 97.0
MONTH: 10 # TMAX: 19.1 # TMIN: 9.6 # TP: 96.2
MONTH: 11 # TMAX: 12.9 # TMIN: 4.1 # TP: 84.5
MONTH: 12 # TMAX: 7.3 # TMIN: -0.4 # TP: 105.2

Weather Forecast Block

format: month/day:hour

DAILY WEATHER DATA:
date;t_min;t_max;t_mean;rh_mean;wind_mean;tp_sum
10/19;17.2;21.4;19.8;78;7.8;0.0
10/20;12.3;18.5;15.5;72;7.2;3.4
10/21;11.1;19.4;14.7;51;4.3;0.0
10/22;13.8;17.8;15.8;57;4.9;2.0
10/23;8.8;14.0;11.4;54;5.5;0.0
10/24;8.9;13.0;10.4;61;4.6;0.0

WEATHER 6-HOUR DATA:
time6h;t_mean;t_min;t_max;rh_mean;wind_mps;wind_dir_deg_mean;precip_mm
10/19:18;20.6;20.4;20.9;73.0;7.3;165.0;0.0
10/19:24;19.4;17.2;21.4;80.0;8.1;167.0;0.0
10/20:06;17.6;17.3;18.5;86.0;8.2;169.0;0.9
10/20:12;15.2;13.2;18.4;80.0;6.8;236.0;2.6
10/20:18;13.6;12.3;15.4;66.0;7.3;253.0;0.0
10/20:24;15.5;15.0;16.2;57.0;6.4;272.0;0.0
10/21:06;13.2;12.3;14.5;55.0;4.6;281.0;0.0
10/21:12;11.5;11.1;12.1;58.0;3.4;260.0;0.0
.....

Figure 3: Schematic representation of the contextual weather data provided to the system as input for hierarchical report generation.

(a)

Weather Summary

During the period from October 19 to October 24 in Cork, Ireland, temperatures exhibited a gradual cooling trend. Initially, daytime maximum temperatures were around 14.4 to 15.5°C, but by October 23 and 24, maxima had decreased to near 10-11.7°C, indicating a transition to cooler autumn conditions. Relative humidity remained generally high, often exceeding 80%, supporting moist conditions typical for this region in mid-autumn. Precipitation was intermittent but light to moderate, with several days recording small rainfall amounts totaling up to 5.4 mm on October 23, consistent with the region's climatological expectations for October. Winds were predominantly moderate, with speeds mostly between 2 and 7.8 m/s, and a general shift from southerly to westerly and northwesterly directions, suggesting passage of a weak frontal system or changing synoptic flow. Overall, the weather pattern reflects a typical autumnal progression with cooling temperatures, moist air, and variable but modest precipitation.

Weather Keywords

`['cooling_trend', 'light_rain', 'moist_conditions', 'frontal_passage', 'autumn_transition']`

Cork, Ireland — 51.903614° N, -8.468399° E

Weather Forecast Block

	time	weather	temp	feels_like	dwp	humidity	wind_deg	wind_speed	wind_gust	precipitation	visibility
1	10/19:21	CLOUDS	14.36	14.12	12.23	87	196	5.03	10.45	0	10000
2	10/19:22	CLOUDS	13.89	13.66	12.11	89	190	4.22	10.66	0	10000
3	10/19:23	RAIN	13.5	13.26	11.89	90	194	4.81	10.54	0.14	10000
4	10/20:00	RAIN	12.87	12.64	11.77	93	174	4.95	10.54	0.43	10000
...	...	...	...	...	...	...	...	...	...	...	...
116	10/24:16	RAIN	10.67	10.04	8.43	86	6	5	7.66	0.24	10000
117	10/24:17	CLOUDS	11	10.3	8.05	82	3	5.2	8.34	0	10000
118	10/24:18	CLOUDS	11.34	10.57	7.64	78	359	5.4	9.81	0	10000
119	10/24:19	CLOUDS	11.67	10.82	7.19	74	356	5.63	9.69	0	10000
120	10/24:20	CLOUDS	10.9	10.06	7.03	77	350	5.39	10.36	0	10000

(b)

Weather Summary

During the period from October 20 to October 25 in Manila, the weather exhibited relatively stable tropical conditions typical for this time of year. Temperatures remained warm with daily maximums generally ranging from 27.6°C to 31.8°C and minimums around 25.4°C to 26.9°C, slightly below the climatological mean maximum of about 31.5°C for October, indicating a mild warm anomaly. Relative humidity was consistently high, averaging between 66% and 90%, supporting a moist atmosphere conducive to light precipitation events. Precipitation was intermittent and light, with daily totals mostly under 5 mm, reflecting scattered showers rather than heavy rainfall. Winds were generally light to moderate, predominantly from easterly to southeasterly directions, with speeds mostly below 4 m/s, indicating stable maritime tropical air mass influence without significant frontal activity or strong wind events.

Weather Keywords

`['light_rain', 'stable_conditions', 'marine_influence', 'warm_anomaly', 'clear_sky']`

Manila, Philippines — 14.5995° N, 120.9842° E

Weather Forecast Block

	time	weather	temp	feels_like	dwp	humidity	wind_deg	wind_speed	wind_gust	precipitation	visibility
1	10/20:03	CLOUDS	26.49	26.49	24.9	91	116	2.66	3.56	0	10000
2	10/20:04	CLOUDS	26.37	26.37	24.6	90	94	3.1	4.33	0	10000
3	10/20:05	CLOUDS	26.22	26.22	24.45	90	98	2.69	4	0	10000
4	10/20:06	CLOUDS	26.21	26.21	24.25	89	92	2.17	3.45	0	10000
...	...	...	...	...	...	...	...	...	...	...	...
116	10/24:22	RAIN	28.47	32.42	23.61	75	105	2.86	3.36	0.17	10000
117	10/24:23	RAIN	27.88	31.23	23.26	76	107	3.51	4.01	0.17	10000
118	10/25:00	CLOUDS	27.56	30.65	23.17	77	107	3.23	3.63	0	10000
119	10/25:01	CLOUDS	27.24	30.07	23.07	78	106	2.97	3.25	0	10000
120	10/25:02	CLOUDS	26.91	29.37	22.75	78	106	2.71	2.86	0	10000

(c)

Weather Summary

Over the period from October 20 to October 24 in Chennai, the weather exhibited a gradual cooling trend accompanied by persistent high humidity and intermittent rainfall. Temperatures peaked around 30.7°C on October 21 but steadily declined to a maximum near 25.3°C by October 24, with minimum temperatures dropping from about 26.6°C to 23.4°C. Relative humidity remained elevated, generally between 73% and 92%, indicating moist atmospheric conditions conducive to cloud formation and precipitation. Wind speeds increased notably from around 3.3 m/s on October 20 to over 10 m/s by October 24, with a shift in wind direction from easterly to more northwesterly, suggesting the influence of a passing weather system or frontal boundary. Precipitation was significant early in the period, with daily totals reaching up to 27.7 mm on October 20, then tapering off towards October 24, reflecting a waning rain event. Overall, the data indicate a transition from a moist, relatively warm and windy period with moderate rainfall to cooler, drier, and windier conditions by the end of the interval.

Weather Keywords

`['heavy_rain', 'frontal_passage', 'strong_wind', 'overcast', 'stable_conditions']`

Chennai, India — 13.0827° N, 80.2707° E

Weather Forecast Block

	time	weather	temp	feels_like	dwp	humidity	wind_deg	wind_speed	wind_gust	precipitation	visibility
1	10/20:00	CLOUDS	27.15	31.17	25.37	90	61	3.73	4.54	0	10000
2	10/20:01	CLOUDS	27.3	31.23	24.95	87	61	3.63	4.44	0	10000
3	10/20:02	CLOUDS	27.42	31.18	24.48	84	61	3.25	4.13	0	10000
4	10/20:03	RAIN	27.13	30.45	24.19	84	69	2.94	3.84	1.09	10000
...	...	...	...	...	...	...	...	...	...	...	...
116	10/24:19	CLOUDS	24.66	25.17	20.15	76	275	6.7	10.11	0	10000
117	10/24:20	CLOUDS	24.96	25.44	20.01	74	270	6.96	9.56	0	10000
118	10/24:21	CLOUDS	25.07	25.51	19.67	72	271	6.42	9.84	0	10000
119	10/24:22	CLOUDS	25.10	25.50	19.33	70	272	5.89	10.12	0	10000
120	10/24:23	CLOUDS	25.29	25.60	19.2	69	274	5.35	10.39	0	10000

Figure 4: Illustrative examples of generated weather reports for three different geographic locations, demonstrating the system's ability to tailor narratives to regional weather conditions.

Reproducibility Checklist

Instructions for Authors:
This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) no

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of

the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) Type your response here

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) Type your response here

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) Type your response here

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) Type your response here

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) Type your response here

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) Type your response here

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) Type your response here

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) Type your response here

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) Type your response here

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) Type your response here

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) Type your response here

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) Type your response here